Resource allocation for cache-enabled cloud-based small cell networks

Article (Accepted Version)

Resource Allocation for Cache-enabled Cloud-based Small Cell Networks

Xiuhua Li, Xiaofei Wang*, Zhengguo Sheng,
Huan Zhou, and Victor C. M. Leung

Abstract

To address the serious challenge of satisfying explosively increasing multimedia content requests from a massive number of users in mobile networks, deploying content caching in base stations to offload network traffic while satisfying content requests locally has been regarded as an effective approach to enhance the network performance. Moreover, content delivery via wireless transmissions in a cache-enabled mobile network needs to be optimized taking the proactive caching policy into consideration. Accordingly, in this paper, we investigate and propose an efficient resource allocation framework for cache-enabled cloud-based small cell networks (C-SCNs) to achieve the benefits of content caching by considering two phases, i.e., content placement and content delivery. In particular, for the content placement phase, we propose a low-complexity distributed popularity-based framework for allocating cache sizes of SBSs to popular contents, in order to offload network traffic and satisfy content requests locally. For the content delivery phase, we propose a low-complexity joint user association and subcarrier-power allocation scheme for min-rate guaranteed content delivery over orthogonal frequency division

This work is based in part on a conference paper presented at IEEE 86th Vehicular Technology Conference (VTC-Fall), 24-27 September 2017, Toronto, Canada. This work is supported in part by a China Scholarship Council Four Year Doctoral Fellowship, the Canadian NSERC through grants RGPIN-2014-06119 and RGPAS-462031-2014, China NSFC (Youth) through grant 61702364.

*Corresponding author.

X. Li and V.C.M. Leung are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada V6T1Z4 (e-mail: {lixiuhua, vleung}@ece.ubc.ca).

X. Wang is with Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin, China 300072 (e-mail: xiaofeiwang@tju.edu.cn).

Z. Sheng is with the Department of Engineering and Design, University of Sussex, Falmer Brighton, United Kingdom BN19RH (e-mail: z.sheng@sussex.ac.uk).

H. Zhou is with College of Computer and Information Technology, China Three Gorges University, Yichang, China 443002 (e-mail: zhouhuan117@gmail.com).

multiple access (OFDMA) based downlink transmissions. Trace-based simulations and numerical results demonstrate the effectiveness of the proposed schemes in the cache-enabled C-SCNs.

Index Terms

Resource allocation, cloud-based small cell network, content caching, traffic load.

I. Introduction

With the growing popularity of smart portable devices such as smartphones and tablets, and online social communities such as Facebook and Twitter, requests for multimedia contents including video, photos, and audio from mobile users are experiencing explosive growth [1]. For mobile network operators (MNOs), satisfying these content requests cost-effectively has become a serious challenge. This problem is further worsened by the scarcity of network resources especially in the radio access networks (RANs) and backhaul networks [2]–[4]. To address the needs to deliver a massive amount of contents with satisfactory Quality of Service (QoS), next generation mobile networking technologies, involving advanced network architectures and new data transmission techniques [5]–[9], are emerging to support the growing network traffic load effectively.

Caching contents at the edges, e.g., base stations (BSs) of mobile networks has recently attracted much attention as an effective approach for offloading network traffic while satisfying QoS, by bringing contents closer to users and then satisfying their content requests locally [1], [6]. There have been a great number of studies focusing on the design of content caching schemes in mobile networks. For instance, cooperative multi-cell caching in [10]–[12] and FemtoCaching in [13] were proposed to cache popular contents in BSs or small BSs (SBSs), aiming at offloading network traffic from massive content downloads and increasing the number of served mobile users. In addition, the concept of Caching-as-a-Service (CaaS) was proposed in [14], focusing on the framework design of virtual caching for offloading network traffic in Cloud-based RANs (C-RANs). Moreover, cooperative BS caching frameworks were proposed in [1], [6], [15], [16], in order to facilitate offloading network traffic by bringing contents closer to users, and improving users' QoS by reducing the time delay of accessing contents. However, most of these studies only focus on content placement, and do not explore the last mile of content delivery via wireless transmissions from BSs to users.

In this paper, we aim to explore the joint resource allocation[1] design of content placement as well as content delivery via wireless transmissions. In practice, popular contents can be stored in BS caches for a long period due to the relatively slow changes of content popularity, while scheduling wireless transmissions of content delivery from BSs to users requires instantaneous channel state information (CSI) of wireless cellular links and is inherently a short-time process. Accordingly, in order to achieve the potential gains of content caching and enhance network capacity, when designing schemes for content delivery via wireless transmissions, we can assume that the states of the caches (i.e., caching status) are static during the wireless transmissions. Furthermore, the corresponding resource allocation is essential in the scheme design. However, there are only a few studies focusing on designing resource allocation schemes for wireless transmissions of content delivery in cache-enabled systems. For instance, in [17], a pricing and resource allocation framework was proposed based on stochastic geometry optimization, aiming to maximize the profit of video caching in small cell networks. The studies in [18], [19] proposed resource allocation schemes for software defined networking, caching and computing, focusing on minimizing the system costs. However, wireless transmissions for content delivery have not been considered in [17]–[19]. In [20], [21], multicast beamforming schemes were proposed for content delivery via wireless transmissions from BSs to users with given caching status. However, the study in [20] only focused on the theoretical analysis of system performance, and did not take into account the detailed scheme design of resource allocation for real-time content delivery satisfying the QoS requirements of users. Besides, the study in [21] did not explore resource allocation with the technique of orthogonal frequency-division multiple access (OFDMA), which has been widely employed in contemporary wireless access networks. In [22], a resource allocation scheme was proposed for minimizing the total transmit power in cache-enabled OFDMA C-RANs, but it did not take into consideration the limit of maximum transmit power of each BS. Hence, resource allocation for content delivery via wireless transmissions in cache-enabled mobile networks with the technique of OFDMA is still not well explored.

To fill the gap by extending our previous work [4], this paper focuses on designing efficient resource allocation frameworks for cache-enabled cloud-based small cell networks (C-SCNs)

[1]Note that in cache-enabled mobile networks, content placement can also be regarded as a kind of resource allocation, where the storage sizes of caches are allocated to contents.

to achieve the potentials of content caching. Two phases, i.e., content placement and content delivery, are considered. Specifically, in the content placement phase, to maximize network traffic offloading while satisfying content requests locally, we propose a low-complexity distributed popularity-based framework for allocating cache sizes of SBSs to popular contents. Wireless transmissions for content delivery from SBSs to users are considered in the content delivery phase, given the caching status in the network. We propose a joint user association and subcarrier-power allocation scheme for min-rate guaranteed content delivery via OFDMA downlink transmissions. Further, to address the complexity of the formulated NP-hard optimization problem regarding wireless content delivery, we use the alternating direction method of multipliers (ADMM) [23]–[25] to split the problem into a set of simpler sub-problems for which optimal solutions can be easily achieved, and propose corresponding methods for solving the sub-problems as well as the whole problem with low complexity. Numerical results from trace-based and Monte Carlo simulations demonstrate the effectiveness of the proposed schemes in the cache-enabled C-SCNs.

The rest of this paper is organized as follows. In Section II, we introduce the network architecture of cache-enabled C-SCNs. In Section III and Section IV, respectively, we propose schemes for content placement based on popularity, and for wireless content delivery. In Section V, we discuss the details of implementing the caching policy. Numerical results from trace-based and Monte Carlo simulation to evaluate the performance of the proposed schemes are shown in Section VI. Finally, Section VII concludes this paper.

II. Network Architecture

A general network architecture of cache-enabled C-SCNs is illustrated in Fig. 1 [4], [6]. Through backhaul links in the C-SCN, a cloud central unit (CCU) is connected to the Enhanced Packet Core (EPC), while the EPC is connected to the Internet. Over the Internet, some service providers (SPs), e.g., Facebook, Netflix, YouTube, offer different kinds of multimedia contents. Besides, in the RAN, some small cells cover the whole area to support various service requests from mobile users via wireless cellular links. Moreover, all the SBSs are connected to the CCU through fronthaul links with low latency and high capacity. Each user in the C-SCN can be associated with and served by multiple SBSs based on the actual channel conditions.

In particular, to offload the massive network traffic caused by downloading of contents from

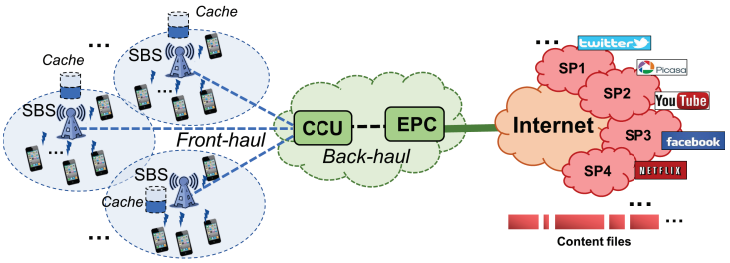


Fig. 1. Illustration of network architecture of cache-enabled C-SNCs.

SPs over the Internet and improve the QoS of users, a limited amount of caches are deployed at all the SBSs, and thus each SBS can cache some contents to bring them closer to mobile users and satisfy as many content requests locally as possible. The CCU makes decisions on how to effectively push popular contents to the caches in its connected SBSs, and maintains a list of the cached contents in all the SBSs at the cost of a small amount of signaling control overhead that is assumed to be negligible [26]. All the required computations to enable this process are performed in the CCU.

Popular contents can be stored in the caches deployed in SBSs for a relatively long period since content popularity generally changes slowly. For instance, short-lifetime popular news with short videos are updated every few hours, while long-lifetime new movies and new music videos are posted weekly and monthly, respectively [10], [13]. However, scheduling wireless transmissions for content delivery from BSs to users requires instantaneous channel state information (CSI) of wireless cellular links and is inherently a short-time process with a time frame of a few minutes. Thus, to achieve the benefits of content caching in the C-SCN, the design of the corresponding scheme can be divided into two phases as follow:

- *Content Placement Phase*: In this phase, the CCU makes decisions on how to store contents in all the SBSs to explore the maximum capacity of the given network infrastructure, by allocating a cache size for each content in each SBS. Due to the relatively slow changes of content popularity, the placement of contents in the C-SCN can remain static over a relatively long time.
- *Content Delivery Phase*: In this phase, given the caching status of all the contents in the C-SCN, in order to dynamically satisfy a content request from a user, the user's associated SBSs either return the content via wireless links with the technique of Coordinated Multi-

Point (CoMP) transmissions if the content is locally available, or route the content request to the CCU. Once a content request routed from an SBS is received at the CCU, the CCU downloads the content directly via backhaul links over the Internet from the respective SP. In particular, content delivery via wireless transmissions from SBSs to users is a process of relatively short time duration in response to the instantaneous content requests from users, and its design involves the allocation of radio resources, e.g., transmit power and bandwidth in each SBS.

In this paper, we consider the case of a cache-enabled C-SCN with a CCU, M single-antenna SBSs (denoted by $\mathcal{M} = \{1, 2, \ldots, M\}$), F popular contents (denoted by $\mathcal{F} = \{1, 2, \ldots, F\}$), and K *active* single-antenna users (denoted by $\mathcal{K} = \{1, 2, \ldots, K\}$) during a time period[2]. We consider that OFDMA is used for wireless transmissions in the C-SCN with N non-overlapping subcarriers (denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$) of the same bandwidth B_s. We assume that the capacities of the fronthaul and backhaul links are sufficiently large to support all the content requests with the employ content caching policy in the C-SCN [26].

Due to the time diversity of the two phases mentioned above, the scheme design in the content placement phase focuses on enhancing the long-time network performances from the perspective of the whole network, while that in the content delivery phase focuses on the short-time wireless transmissions for improving the QoS from the perspective of a specific group of users. In the following, we present the scheme designs for the above two phases of content caching involving resource allocation for the cache-enabled C-SCN in Sections III and IV, respectively.

III. Content Placement Based on Popularity

In this section, we introduce the model of content placement in the considered cache-enabled C-SCN, and propose a distributed content placement framework.

A. Content Placement Model

In the content placement phase, each content f is assumed to be either entirely cached in SBS_m or not, respectively denoted by $x_m^f = 1$ or $x_m^f = 0$. From a practical perspective, we assume that

[2]Note that in practice the total number of users is much greater than K in the network. We only consider the short-time wireless transmissions during content delivery to satisfy the content requests from a given number of users that are active during a specific time period, e.g., a time slot.

different contents have different storage sizes, which are denoted by $\{s_1, s_2, \ldots, s_F\}$. Denote the cache storage sizes of SBSs as $\{S_1, S_2, \ldots, S_M\}$. We denote ϕ_m as the average overall arrival rate of content requests received at SBS$_m$, which can be defined as the ratio of the total number of content requests received at SBS$_m$ to the entire time period considered, In a similar way, we define the average arrival rate of the requests for content o_f received at SBS$_m$ as ϕ_m^f.

Particularly, based on [10], we also use Mandelbrot-Zipf (MZipf) distribution to model the global popularity of the contents in the network, denoted as $\{P_1, P_2, \ldots, P_F\}$, which can be expressed as

$$P_f = \frac{(\gamma_f + c_0)^{-\beta}}{\sum_{i \in \mathcal{F}} (\gamma_i + c_0)^{-\beta}}, \ \forall\, f \in \mathcal{F} \tag{1}$$

where γ_f denotes the rank of the content f in the descending order of global content popularity, while $c_0 \geq 0$ and $\beta > 0$ denote the plateau factor and the skewness factor, respectively. Note that if the plateau factor c_0 takes the value of zero, then the MZipf distribution reduces to Zipf distribution [13]. Besides, denote ρ_m^f as the local popularity of content f in the m-th small cell, which is defined as the ratio of the number of requests of content f in the m-th small cell to the total number of requests of all the contents in the network. As a result, we have $P_f = \sum_{m \in \mathcal{M}} \rho_m^f$ and $\phi_m^f = \frac{\rho_m^f}{\sum_{i \in \mathcal{F}} \rho_m^i} \phi_m$ for $\forall m \in \mathcal{M}, \forall f \in \mathcal{F}$. We assume that the global/local popularity of each content can be determined in advance or predicted by the system through learning and analysis of user behavior and preference [1], and thus it is available in the network.

B. Distributed Content Placement Framework

In this phase, our objective is to minimize the expected sum of traffic load caused by utilizing backhaul/fronthaul links for downloading contents directly from SPs to SBSs via CCU. This optimization problem is equivalent to maximizing the expected sum of traffic offload among SBSs while satisfying content requests locally. Here, similar to [10], we also regard $\phi_m^f s_f$ as the generated average traffic load for the requests of content f received at SBS$_m$.

Thus, under the constraints of limited cache storage capacity of SBSs, the corresponding optimization problem for the popularity-based content placement can be formulated as

$$\min_{\{x_m^f\}} \ \sum_{m \in \mathcal{M}} \sum_{f \in \mathcal{F}} x_m^f \phi_m^f s_f \tag{2a}$$

$$\textit{s.t.} \ \sum_{f \in \mathcal{F}} x_m^f s_f \leq S_m, \quad \forall m \in \mathcal{M}, \tag{2b}$$

$$x_m^f \in \{0,1\},\ \forall m \in \mathcal{M}, \forall f \in \mathcal{F}. \tag{2c}$$

The above problem in (2) can be further separated into M independent single knapsack problems and solved in a distributed manner with the greedy method [1], thereby achieving a distributed content placement framework for the cache-enabled C-SCN.

IV. Content Delivery via Wireless Transmissions

In this section, we introduce the model of wireless transmissions for content delivery in the considered cache-enabled C-SCN, and propose an effective framework based on ADMM.

A. Wireless Transmission Model

In the content delivery phase, we consider the downlink OFDMA transmissions of a cache-enabled C-SCN. We assume that interference between adjacent cells can be avoided even if the maximum subcarrier reuse factor of 1 is applied. In addition, we assume that the CCU performs a centralized control of content delivery taking available CSI and the information of users' content requests into account. The downlink channel is assumed to be slotted, and the scheme design of resource allocation is performed on a slot-by-slot basis over much shorter time intervals than those of status changes in content placement, which are relatively static and known to the CCU during the process of content delivery.

Recall the indictors $x_m^f \in \{0,1\}$ for whether SBS$_m$ caches content f or not, which is available with the proposed content placement scheme. Denote $y_k^f \in \{0,1\}$ as the indictor for whether content f is requested by user k or not, and each user accesses only one content in a time slot, i.e., $\sum_{f=1}^{F} y_k^f = 1, \forall k \in \mathcal{K}$. Denote $\mathcal{S}_k = \{m | x_m^f = y_k^f = 1, m \in \mathcal{M}, f \in \mathcal{F}\}, k \in \mathcal{K}$ as the set of SBSs that cache the requested content of user k, $\mathcal{K}_1 = \{k | \mathcal{S}_k \neq \emptyset, k \in \mathcal{K}\}$ as the set of users whose requested contents are locally available, and $\mathcal{K}_0 = \mathcal{K} \setminus \mathcal{K}_1$ as the set of users whose requested contents are not cached and need to be downloaded from SPS over the Internet through backhaul links. In order to offload network traffic and reduce network costs with the help of the given content placement scheme, each user in the set $\mathcal{K}_1$ is required to be associated with at least one of the SBSs that cache the requested content while each user in the set $\mathcal{K}_0$ can be associated with any SBS in the network. Denote $h_{k,m,n}$ and $p_{k,m,n}$, respectively, as the complex channel gain that takes into account of both large-scale and small-scale fading and transmit power from SBS$_m$

to user k on subcarrier n. Denote $\delta_{k,n} \in \{0,1\}$ and $\delta_{k,m,n} \in \{0,1\}$ as the respective indictors for whether or not subcarrier n is allocated to user k in the network and from SBS$_m$, respectively. Here, $\{\delta_{k,n}\}$ satisfies $\sum_{k\in\mathcal{K}} \delta_{k,n} \leq 1, \forall n \in \mathcal{N}$, while $\delta_{k,m,n} = 0, \forall k \in \mathcal{K}_1, \forall m \in \mathcal{M}\backslash\mathcal{S}_k, \forall n \in \mathcal{N}$ holds[3]. Physically, if and only if the transmit power $p_{k,m,n} > 0$ holds can subcarrier n be regarded as being allocated to user k by SBS$_m$, and user k is associated with SBS$_m$ on subcarrier n. Thus, according to our previous work [25], we can derive the mathematical relationship between the subcarrier allocation $\delta_{k,m,n}$ and $p_{k,m,n}$ as

$$\delta_{k,m,n} = \text{sign}(p_{k,m,n}) \text{ and } \delta_{k,m,n} p_{k,m,n} = p_{k,m,n} \tag{3}$$

where $\text{sign}(x) \triangleq \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x = 0, \end{cases}$ $(x \geq 0)$ is the step function. In a similar way, we can also derive the relationship between $\delta_{k,n}$ and $\delta_{k,m,n}$ as

$$\delta_{k,n} = \max_{m\in\mathcal{M}} \{\delta_{k,m,n}\} = \text{sign}\big(\sum_{m\in\mathcal{M}} p_{k,m,n}\big). \tag{4}$$

Thus, based on (3) and (4), we can conclude that the joint user association and subcarrier-power allocation problem can be transformed into an equivalent power allocation problem by introducing the defined step function.

Moreover, based on (3) and by allowing multiple SBSs to transmit to one user by Coordinated Multi-Point (CoMP) transmissions, e.g., employing the maximum ratio transmission (MRT) technique [25], we can get the signal-to-noise ratio (SNR) of user k on subcarrier n from all the SBSs as

$$\rho_{k,n} = \frac{\sum_{m\in\mathcal{M}} p_{k,m,n} \mid h_{k,m,n} \mid^2}{\sigma_N^2}, \ \forall k \in \mathcal{K}, \forall n \in \mathcal{N} \tag{5}$$

where σ_N^2 is the power of the zero-mean additive white Gaussian noise (AWGN) at the receiver input. Thus, the channel capacity of user k on subcarrier n from all the SBSs can be expressed as

$$r_{k,n} = B_s \log_2(1+\rho_{k,n}), \forall k \in \mathcal{K}, \forall n \in \mathcal{N}. \tag{6}$$

Then we can get the overall data rate of user k from all the SBSs and subcarriers as

$$R_k = \sum_{n\in\mathcal{N}} r_{k,n}, \quad \forall k \in \mathcal{K}. \tag{7}$$

[3]Note that for some (k,n), $\sum_{m\in\mathcal{M}} \delta_{k,m,n} > 1$ may hold in the resource allocation scheme. In other words, each subcarrier can be allocated to at most one user to avoid interference, but multiple SBSs can allocate the same subcarrier to the same user in the network.

B. Problem Formulation for Content Delivery

In this phase, our objective is to maximize the weighted sum of data rates of all the users in a cache-enabled C-SCN based on joint user association and subcarrier-power allocation for the OFDMA downlink transmissions of min-rate guaranteed content delivery. The overall optimization problem for content delivery via wireless transmissions is formulated as

$$\max_{\mathbf{P}\in\mathbb{R}^{K\times M\times N}} \lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k + \sum_{k\in\mathcal{K}_0} \omega_k R_k \tag{8a}$$

$$\textit{s.t.}\ p_{k,m,n} \geq 0, \quad \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{8b}$$

$$\sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} p_{k,m,n} \leq p_m^{\max}, \quad \forall m \in \mathcal{M}, \tag{8c}$$

$$R_k \geq C_k^{\min}, \forall k \in \mathcal{K}, \tag{8d}$$

$$\delta_{k,m,n} = \text{sign}(p_{k,m,n}), \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{8e}$$

$$\delta_{k,m,n} = 0, p_{k,m,n} = 0, \forall k \in \mathcal{K}_1, \forall m \in \mathcal{M} \setminus \mathcal{S}_k, \forall n \in \mathcal{N}, \tag{8f}$$

$$\delta_{k,n} = \max_{m\in\mathcal{M}} \{\delta_{k,m,n}\}, \quad \forall\, k \in \mathcal{K}, \forall\, n \in \mathcal{N}, \tag{8g}$$

$$\sum_{k\in\mathcal{K}} \delta_{k,n} \leq 1, \quad \forall\, n \in \mathcal{N} \tag{8h}$$

where the weighting factors $\lambda \geq 1$ and ω_k denote the network priority of the users whose requested contents are locally cached and the individual priority of user k, respectively; $\mathbf{P} = \{p_{k,m,n}\}^{K\times M\times N}$; $p_m^{\max}$ denotes the maximum transmit power of SBS$_m$; $C_k^{\min}$ denotes the required minimum data rate of user k. Note that all the sets $\mathcal{K}_0$, $\mathcal{K}_1$ and $\{\mathcal{S}_k\}$ are dependent on both the achieved content placement policy $\{x_m^f\}$ and the users' content requests $\{y_k^f\}$, and can be pre-determined before content delivery via wireless transmissions. Mathematically, the transmit power constraints in (8b) and (8c) imply that

$$0 \leq p_{k,m,n} \leq p_m^{\max}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{9}$$

which is useful in the algorithm design [27]. Denote $\mathcal{P}$ as the feasible solution set of the problem in (8). Clearly, the problem in (8) is a mixed 0-1 nonconvex optimization problem and thus is NP-hard [28].

C. ADMM-based Decomposition

In order to solve the formulated NP-hard optimization problem in (8), we aim at providing a suboptimal solution with low-complexity by employing the ADMM used in [23]–[25]. The main idea of ADMM is to decompose the complex original problem into a series of subproblems that are much simpler to solve, and to combine the solutions to the subproblems together in a principled manner to obtain the solution to the original problem finally. Accordingly, based on the idea of ADMM, we first divide the large-scale constraints in (8) into two small-scale groups and define two sets as

$$\mathcal{S}_P = \left\{ \mathbf{P} \in \mathbb{R}^{K\times M\times N} \middle| \ (8c), (8d), (8f) \text{ and } (9) \right\}, \tag{10}$$

$$\mathcal{S}_Q = \left\{ \mathbf{P} \in \mathbb{R}^{K\times M\times N} \middle| \ (8e) - (8h) \text{ and } (9) \right\}. \tag{11}$$

Clearly, the set $\mathcal{S}_P$ aims to satisfy the constraints of the required minimum data rates of the users and the maximum transmit power of the SBSs, and is convex. The set $\mathcal{S}_Q$ is to satisfy the constraints of user association and subcarrier allocation, but is discrete and nonconvex. Consequently, the feasible solution set satisfies $\mathcal{P} = \mathcal{S}_P \cap \mathcal{S}_Q$, i.e., $\mathbf{P} \in \mathcal{S}_P$ and $\mathbf{P} \in \mathcal{S}_Q$. Then the problem in (8) can be rewritten as

$$\min_{\mathbf{P}\in\mathcal{S}_\mathbf{P}, \mathbf{Q}\in\mathcal{S}_\mathbf{Q}} \ F(\mathbf{P}) \tag{12a}$$

$$s.t. \ \mathbf{P} = \mathbf{Q} \tag{12b}$$

where $F(\mathbf{P}) = -\lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k - \sum_{k\in\mathcal{K}_0} \omega_k R_k$, and $\mathbf{Q} = \{q_{k,m,n}\}^{K\times M\times N} \in \mathbb{R}^{K\times M\times N}$ is an introduced variable matrix. Thus, the problem in (8) is equivalently transformed to the problem in (12) with an equality constraint.

Then the problem in (12) can be turned into a minimization problem by introducing the corresponding augmented Lagrangian function as

$$\mathcal{L}(\mathbf{P}, \mathbf{Q}, \mathbf{L}, \theta) = F(\mathbf{P}) + \langle \mathbf{P} - \mathbf{Q}, \mathbf{L} \rangle + \frac{\theta}{2}(\|\mathbf{P} - \mathbf{Q}\|_2^2) \tag{13}$$

where $\mathbf{L} \in \mathbb{R}^{K\times M\times N}$ denotes the Lagrange multiplier matrix associated with the constraint (12b); $\theta > 0$ is a quadratic penalty scalar; $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the sum of all the elements of $\mathbf{x} \circ \mathbf{y}$, where $\circ$ denotes the Hadamard product.

By employing ADMM-based decomposition, the joint optimization problem w.r.t. the augmented Lagrangian function in (13) can be decomposed into three subproblems as follow:

1) Subproblem 1: Optimization of $\mathbf{P}$ with fixed $\mathbf{Q}$, $\mathbf{L}$ and θ, which is formulated as

$$\min_{\mathbf{P}\in\mathcal{S}_\mathbf{P}} F(\mathbf{P}) + \frac{\theta}{2}\|\mathbf{P}-\mathbf{C}_P\|_2^2 \tag{14}$$

where $\mathbf{C}_P = \mathbf{Q} - \frac{\mathbf{L}}{\theta}$ is a constant matrix w.r.t. $\mathbf{P}$.

2) Subproblem 2: Optimization of $\mathbf{Q}$ with fixed $\mathbf{L}$ and θ, which is formulated as

$$\min_{\mathbf{Q}\in\mathcal{S}_\mathbf{Q}} \|\mathbf{Q}-\mathbf{C}_Q\|_2^2 \tag{15}$$

where $\mathbf{C}_Q \triangleq \mathbf{P}^* + \frac{\mathbf{L}}{\theta}$ is a constant matrix w.r.t. $\mathbf{Q}$, and $\mathbf{P}^*$ is the optimal solution to Subproblem 1.

3) Subproblem 3: Updating of $\mathbf{L}$ and θ iteratively with the achieved $(\mathbf{P}^*, \mathbf{Q}^*)$, where $\mathbf{Q}^*$ is the optimal solution to Subproblem 2.

In the above ADMM-based decomposition, Subproblem 1 is a convex optimization problem, and thus its optimal solution can be easily achieved by employing optimization techniques (e.g., subgradient method) as shown in Section IV-D. Subproblem 2 is a discrete nonconvex optimization problem, but its optimal solution can be solved with a low-complexity distributed search algorithm as shown in Section IV-E. Subproblem 3 can be solved by updating $\mathbf{L}$ and θ based on the principles of ADMM, and thus the whole original problem in (12) can also be solved as shown in Section IV-F.

In other words, by employing ADMM, the original NP-hard optimization problem is decomposed into a series of simpler subproblems where their optimal solutions can be obtained easily. Note that due to the non-convexity of the original complex problem, the final solution will be suboptimal.

D. Solutions to Subproblem 1

Subproblem 1 in (14) can be rewritten in a standard form as

$$\min_{\mathbf{P}\in\mathbb{R}^{K\times M\times N}} \; -\lambda\sum_{k\in\mathcal{K}_1}\omega_k R_k(\mathbf{P}) - \sum_{k\in\mathcal{K}_0}\omega_k R_k(\mathbf{P}) + \frac{\theta}{2}\|\mathbf{P}-\mathbf{C}_P\|_2^2 \tag{16a}$$

$$\textit{s.t. } 0 \le p_{k,m,n} \le p_m^{\max}, \forall k\in\mathcal{K}, \forall m\in\mathcal{M}, \forall n\in\mathcal{N}, \tag{16b}$$

$$p_{k,m,n} = 0, \;\; \forall k\in\mathcal{K}_1, \forall m\in\mathcal{M}\setminus\mathcal{S}_k, \forall n\in\mathcal{N}, \tag{16c}$$

$$\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} p_{k,m,n} \le p_m^{\max}, \quad \forall m\in\mathcal{M}, \tag{16d}$$

$$- R_k(\mathbf{P}) + C_k^{\min} \leq 0, \quad \forall\, k \in \mathcal{K}, \tag{16e}$$

which is a convex optimization problem w.r.t. $\mathbf{P}$.

Then by employing standard optimization techniques in [29], we can get the corresponding Lagrangian function as

$$\begin{aligned}\mathcal{L}_1(\mathbf{P}, \boldsymbol{\nu}, \boldsymbol{\mu}) = \frac{\theta}{2}\|\mathbf{P}-\mathbf{C}_P\|_2^2 - \sum_{k\in\mathcal{K}_1} (\lambda\,\omega_k + \mu_k) R_k(\mathbf{P}) - \sum_{k\in\mathcal{K}_0} (\omega_k + \mu_k) R_k(\mathbf{P}) \\ + \sum_{k\in\mathcal{K}} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \nu_m p_{k,m,n} - \sum_{m\in\mathcal{M}} \nu_m p_m^{\max} + \sum_{k\in\mathcal{K}} \mu_k C_k^{\min}\end{aligned} \tag{17}$$

where $\boldsymbol{\nu}$ is the Lagrange multiplier vector associated with the constraint (16d) with elements $\nu_m, \forall m \in \mathcal{M}$, and $\boldsymbol{\mu}$ is the Lagrange multiplier vector associated with the constraint (16e) with elements $\mu_k, \forall k \in \mathcal{K}$.

After differentiating $\mathcal{L}_1(\mathbf{P}, \boldsymbol{\nu}, \boldsymbol{\mu})$ w.r.t. $\mathbf{P}$, we can obtain

$$\frac{\partial \mathcal{L}_1}{\partial p_{k,m,n}} = \theta[p_{k,m,n} - (\mathbf{C}_P)_{k,m,n}] + \nu_m - \frac{B_s\,\varpi_k}{\ln 2} \frac{H_{k,m,n}}{1+\sum_{j\in\mathcal{M}} H_{k,j,n}\, p_{k,j,n}}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N} \tag{18}$$

where $H_{k,m,n} = \frac{|h_{k,m,n}|^2}{\sigma_N^2}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}$, and

$$\varpi_k = \begin{cases} \lambda\,\omega_k + \mu_k, \text{ if } k \in \mathcal{K}_1, \\ \omega_k + \mu_k, \text{ if } k \in \mathcal{K}_0. \end{cases} \tag{19}$$

We use the subgradient method to get the optimal solution to Subproblem 1. The multipliers $(\boldsymbol{\nu}, \boldsymbol{\mu})$ are updated in each step as

$$\nu_m^{(t+1)} = \Big[\nu_m^{(t)} + \xi^{(t)}\Big(\sum_{k\in\mathcal{K}}\sum_{n\in\mathcal{N}} p_{k,m,n} - p_m^{\max}\Big)\Big]^+, \forall m \in \mathcal{M}, \tag{20}$$

$$\mu_k^{(t+1)} = \big[\mu_k^{(t)} + \xi^{(t)}\big(C_k^{\min} - R_k^{(t)}\big)\big]^+, \ \forall k \in \mathcal{K} \tag{21}$$

where t is the iteration index, $\xi^{(t)} > 0$ is a step size in the t-th iteration, and $[x]^+ \triangleq \max\{x, 0\}$. If the step sizes $\{\xi^{(t)}\}$ are selected to be sufficiently small, e.g., $\xi^{(t)} = \frac{1+K}{t+K}$, the convergence to the optimal multipliers $(\boldsymbol{\nu}^*, \boldsymbol{\mu}^*)$ with the subgradient method can be guaranteed [29].

In addition, with fixed $(\boldsymbol{\nu}^{(t)}, \boldsymbol{\mu}^{(t)})$, the elements of $\mathbf{P}^*$ are updated as

$$p_{k,m,n}^{(t)} = \begin{cases} 0, \text{ if } k \in \mathcal{K}_1, m \in \mathcal{M} \setminus \mathcal{S}_k, n \in \mathcal{N}, \\ \min\big\{\big[T_{k,m,n}^{(t)}\big]^+, p_m^{\max}\big\}, \text{ otherwise} \end{cases} \tag{22}$$

where $T_{k,m,n}^{(t)} = \frac{b_{k,m,n}^{(t)} + \sqrt{[b_{k,m,n}^{(t)}]^2 + 4H_{k,m,n}a_{k,m,n}^{(t)}}}{2H_{k,m,n}}$, $b_{k,m,n}^{(t)} = c_{k,m,n}^{(t)} H_{k,m,n} - \sum_{j\in\mathcal{M}\setminus\{m\}} H_{k,j,n} p_{k,j,n}^{(t)} - 1$, $a_{k,m,n}^{(t)} = c_{k,m,n}^{(t)}[\sum_{j\in\mathcal{M}\setminus\{m\}} H_{k,j,n} p_{k,j,n}^{(t)} + 1] + \frac{B_s \varpi_k^{(t)}}{\theta \ln 2} H_{k,m,n}$, and $c_{k,m,n}^{(t)} = (\mathbf{C}_P)_{k,m,n} - \frac{\nu_m^{(t)}}{\theta}$.

The procedure of the proposed subgradient method for solving Subproblem 1 is shown in Algorithm 1, and the corresponding complexity and convergence analysis can be found in [29]. Algorithm 1 consists of an inner loop and an outer loop. With the given Lagrangian multipliers $(\boldsymbol{\nu}, \boldsymbol{\mu})$ at each iteration, the inner loop aims to update $\mathbf{P}$, which converges to the unique optimal solution as a result of the convexity of (17). mThe solution $\mathbf{P}$ obtained using Algorithm 1 is optimal to Subproblem 1.

Algorithm 1 Subgradient Algorithm for Solving Subproblem 1 w.r.t. $\mathbf{P}$.

1: **Input:** B_s, σ_N^2, $(h_{k,m,n})_{K\times M\times N}$, λ, $(\omega_k)_{K\times 1}$, $(p_m^{\max})_{M\times 1}$, $(C_k^{\min})_{K\times 1}$, $\mathbf{C}_P$, θ, $\mathbf{P}_{ini}$.
2: Initialize $t = 0$, $\boldsymbol{\nu}^{(0)} \succ 0_{M\times 1}$, $\boldsymbol{\mu}^{(0)} \succ 0_{K\times 1}$, convergence precision $\varrho = 10^{-4}$, maximum iterations $N_{\max} = 20$, $\mathbf{P}^{(0)} = \mathbf{P}_{ini}$.
3: **while** $(\boldsymbol{\nu}, \boldsymbol{\mu})$ not converge **do**
4: **while** not exceed $N_{\max}$ or $\mathbf{P}$ not converge **do**
5: Update $\mathbf{P}^{(t)}$ according to (22).
6: **end while**
7: Set $t \leftarrow t + 1$.
8: Update $\boldsymbol{\nu}^{(t)}$ and $\boldsymbol{\mu}^{(t)}$ based on (20) and (21), respectively.
9: Check the convergence condition: $\|\boldsymbol{\nu}^{(t)} - \boldsymbol{\nu}^{(t-1)}\|_\infty \le \varrho$ and $\|\boldsymbol{\mu}^{(t)} - \boldsymbol{\mu}^{(t-1)}\|_\infty \le \varrho$.
10: **end while**
11: **Output:** $\mathbf{P}$.

E. Solutions to Subproblem 2

Subproblem 2 in (15) can be rewritten as

$$\min_{\mathbf{Q}\in\mathbb{R}^{K\times M\times N}} \|\mathbf{Q} - \mathbf{C}_Q\|_2^2 \tag{23a}$$

$$\textit{s.t. } 0 \le q_{k,m,n} \le p_m^{\max}, \forall k \in \mathcal{K}, \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{23b}$$

$$q_{k,m,n} = 0, \ \forall k \in \mathcal{K}_1, \forall m \in \mathcal{M} \setminus \mathcal{S}_k, \forall n \in \mathcal{N}, \tag{23c}$$

$$\sum_{k\in\mathcal{K}} \max_{m\in\mathcal{M}} \{\text{sign}(q_{k,m,n})\} \le 1, \quad \forall\, n \in \mathcal{N}. \tag{23d}$$

Subproblem 2 is nonconvex, but can be further divided into N subproblems and solved in parallel. Accordingly, we propose a distributed search method in a closed form as shown in Algorithm 2 to find the optimal solution to Subproblem 2. The time complexity of Algorithm 2 is linear, i.e., $O(KMN)$.

Algorithm 2 Distributed Search Algorithm for Solving Subproblem 2 w.r.t. $\mathbf{Q}$.

Input: $\mathbf{P}$, $\mathbf{L}$, θ, $(p_m^{\max})_{M\times 1}$.
Initialize $\mathbf{Q} = \{q_{k,m,n}\}^{K\times M\times N} = \mathbf{0}_{K\times M\times N}$, $\mathbf{T} = \{t_{k,m,n}\}^{K\times M\times N} = \mathbf{0}_{K\times M\times N}$.
Calculate $\mathbf{C}_Q$.
for $n = 1$ to N **do**
  for $k = 1$ to K **do**
    if $k \in \mathcal{K}_1$ **then**
      Set $t_{k,m,n} = \min\{\big[\big(\mathbf{C}_Q\big)_{k,m,n}\big]^+, p_m^{\max}\}$ for $\forall m \in \mathcal{S}_k$.
    else
      Set $t_{k,m,n} = \min\{\big[\big(\mathbf{C}_Q\big)_{k,m,n}\big]^+, p_m^{\max}\}$ for $\forall m \in \mathcal{M}$.
    end if
  end for
  Given n, find $k_n^* = \arg\min_{k\in\mathcal{K}}\{\sum_{m=1}^{M}\big(t_{k,m,n} - \mathbf{C}_Q\big)_{k,m,n}\big)^2\}$.
  if k_n^* multiple **then**
    Select one randomly.
  end if
  Set $q_{k_n^*,m,n} \leftarrow t_{k_n^*,m,n}$ for $\forall m \in \mathcal{M}$.
end for
Output: $\mathbf{Q}$.

F. Solutions to Subproblem 3

After solving Subproblem 1 and Subproblem 2 to find their optimal solutions w.r.t. $\mathbf{P}$ and $\mathbf{Q}$, respectively, Subproblem 3 is concerned with updating the multiplier $\mathbf{L}$ and the quadratic penalty scalar θ. Based on the rules of ADMM, they are updated in each step as

$$\mathbf{L}^{(\tau+1)} = \mathbf{L}^{(\tau)} + \theta^{(\tau)}\big(\mathbf{P}^{(\tau+1)} - \mathbf{Q}^{(\tau+1)}\big), \tag{24}$$

$$\theta^{(\tau+1)} = \min\big\{\theta^{\max}, \Delta\cdot\theta^{(\tau)}\big\} \tag{25}$$

where τ is the iteration index, while $\theta^{\max} > 0$ and $\Delta > 1$ are given positive scalars. The procedure of the proposed ADMM for solving the whole problem in (12) is shown in Algorithm 3. In the iterative process, with the updated multiplier $\mathbf{L}$ and quadratic penalty scalar θ, Algorithm 3 solves Subproblem 1 and Subproblem 2 to update $\mathbf{P}$ and $\mathbf{Q}$, respectively. Moreover, the ADMM converges to the corresponding suboptimal solution to the optimization problem in (12) [23], [30].

Algorithm 3 ADMM for Solving the whole Problem in (12).

1: **Input:** B_s, σ_N^2, $(h_{k,m,n})_{K\times M\times N}$, λ, $(\omega_k)_{K\times 1}$, $(p_m^{\max})_{M\times 1}$, $(C_k^{\min})_{K\times 1}$.
2: Initialize $\tau = 0$, $\mathbf{P}^{(0)} = \mathbf{0}_{K\times M\times N}$, $\mathbf{Q}^{(0)} = \mathbf{0}_{K\times M\times N}$, $\mathbf{L}^{(0)} \succ 0_{K\times M\times N}$, $\theta^{(0)} > 0$, $\theta^{\max} > 0$, $\Delta > 1$, convergence precision $\varepsilon = 10^{-4}$.
3: **while** not converge **do**
4: Update $\mathbf{P}^{(\tau)}$ by solving Subproblem 1.
5: Update $\mathbf{Q}^{(\tau)}$ by solving Subproblem 2.
6: Set $\tau \leftarrow \tau + 1$.
7: Update $\mathbf{L}^{(\tau)}$ and $\theta^{(\tau)}$ according to (24) and (25), respectively.
8: Check the convergence condition: $\|\mathbf{P}^{(\tau)} - \mathbf{Q}^{(\tau)}\|_\infty \leq \varepsilon$.
9: **end while**
10: **Output: P**.

In particular, to initialize Algorithm 3, the multiplier $\mathbf{L}$ can be set randomly. The scalar θ is generally initialized with a small value, e.g., $\theta^{(0)} = 10^{-3}$, while the scalar $\theta^{\max}$ can be initialized as a relatively large value, e.g., $\theta^{\max} = 10^6$. The scalar Δ needs to be initialized properly according to the convergence rate of ADMM, which is neither too large nor too small, e.g., $\Delta = 1.2$. However, with different feasible initializations, ADMM can always converge but needs different numbers of iterations for satisfying the given convergence condition, and its corresponding complexity and convergence analysis can be found in [23], [30], [31].

V. Implementation of Caching Policy

Based on the proposed schemes of content placement and content delivery, the caching policy can be implemented in the CCU management process as shown in Fig. 2. Generally, after collecting and analyzing the global information including user characteristics, content features and communication scenario from SBSs and mobile users, the CCU computes the correspond

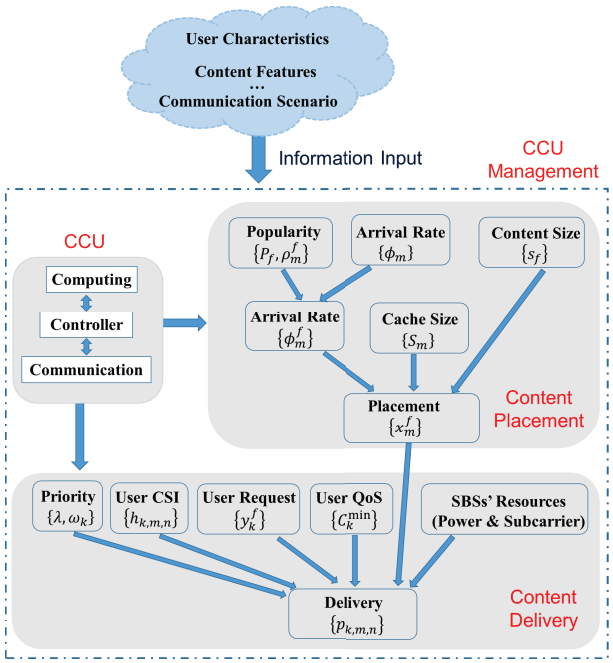


Fig. 2. The CCU management process implementing the caching policy.

caching policy in terms of content placement and content delivery, and communicates with the connected SBSs.

Particularly, in the content placement phase, the global/local popularity $\{P_f, \rho_m^f\}$ and the arrival rate $\{\phi_m\}$ can be obtained to calculate the arrival rate $\{\phi_m^f\}$. Then by considering the arrival rate $\{\phi_m\}$, content sizes $\{s_f\}$ and the cache sizes $\{S_m\}$, the content placement problem can be formulated as in (2). After solving it, the content placement policy $\{x_m^f\}$ can be found.

Moreover, in the wireless transmissions of content delivery phase, the content delivery problem can be formulated as in (8) by considering the obtained content placement policy $\{x_m^f\}$ with the priority $\{\lambda, \omega_k\}$, user CSI $\{h_{k,m,n}\}$, user requests $\{y_k^f\}$, user QoS requirements $\{C_k^{\min}\}$ and SBSs' resources, i.e., maximum transmit power and subcarriers. After solving this problem with the proposed ADMM, the content delivery policy $\{p_{k,m,n}\}$ can be determined.

VI. Evaluation Results

In this section, we evaluate the performance of our proposed schemes of resource allocation in the cache-enabled C-SCN with respect to the considered two phases of content caching. For simulation purposes, the whole service area is set as a circle with a radius of 500 meters, and

fully covered by five small cells, i.e., $M = 5$. The five SBSs, each with a coverage radius of 250 meters, are uniformly distributed in the circle. We further assume that the SBSs have the same cache size and maximum transmit power, i.e., $S_m \equiv S$ and $p_m^{\max} \equiv p^{\max}, \forall\, m \in \mathcal{M}$, respectively.

A. Trace-based Results in Content Placement

In this subsection, we evaluate the performance of our proposed content placement scheme in the cache-enabled C-SCN. We use the trace of a real-world proxy caching system, IRCache as used in [6]. For our simulations, the trace data for 7 days in June 2013 were collected to obtain user requests of popular contents over the Internet as well as their content sizes. The data set consists of $50,000$ popular contents and $4,928$ users, corresponding to $516,135$ content requests. In addition, we use random settings for mapping the association between users and SBSs for calculating the local/global popularity of contents as well as the average arrival rate of content requests.

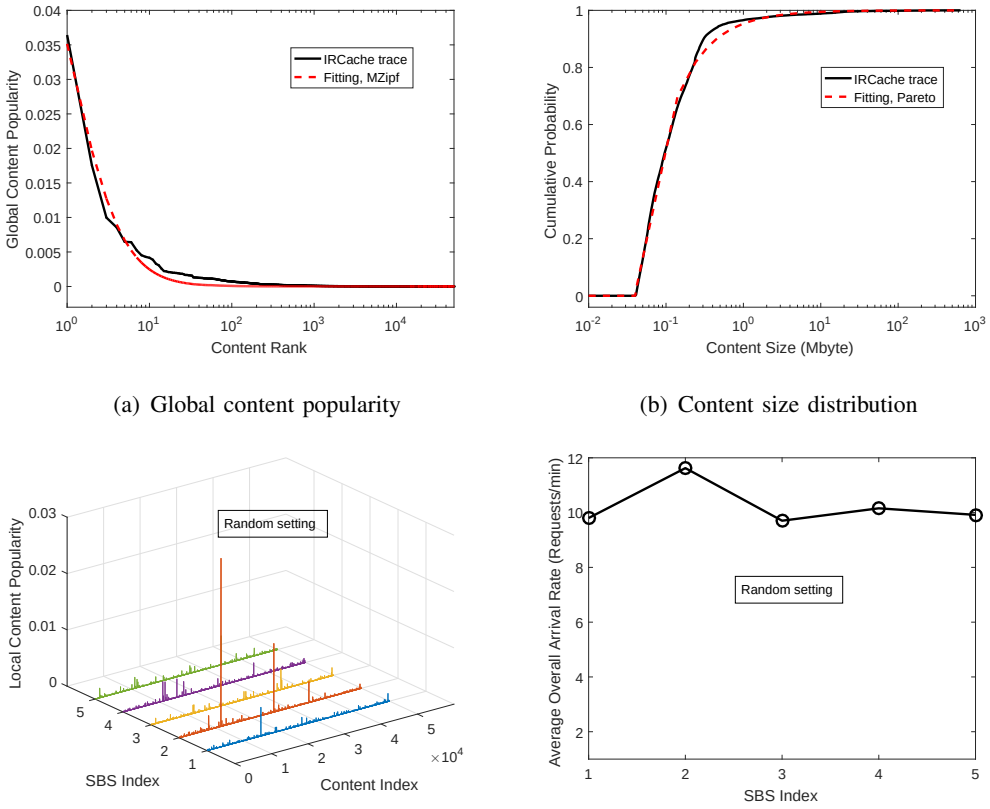
1) Distributions

Fig. 3 shows different distributions in the IRCache trace and random settings. From Fig. 3(a) and Fig. 3(b), we can observe that the actual global content popularity and content size in the IRCache trace can be well fitted by a MZipf distribution and a Pareto distribution, respectively, agreeing with the model on the global content popularity used in (1) and the corresponding conclusion in [33]. Fig. 3(c) and Fig. 3(d) show the local content popularity and average overall arrival rate (around 10 requests per minute) in each SBS, respectively, in the random setting as a result of the joint consideration of the IRCache trace and random user association. Note that the following evaluation results are based on the above practical trace and random setting.

2) Effects of Different Cache Sizes of Each SBS

In particular, we compare the proposed scheme with two baseline schemes as: i) *Most Popular Caching*, derived by caching most popular contents in each SBS with given content popularity, which aims to maximize the cache hit ratio[4] based on the the constraints of cache sizes of SBSs; ii) *Least Recently Used* (LRU) Caching, an online scheme derived from [34] to address the problem of content placement in this paper; iii) *Random Caching*, derived by randomly filling contents in each SBS until the cache is full without any information of content popularity.

[4]The cache hit ratio is defined as the ratio of tsupported number of content requests by the caching scheme to the total number of content requests in the network, i.e., $\big(\sum_{m\in\mathcal{M}}\sum_{f\in\mathcal{F}} x_m^f \phi_m^f\big)/\big(\sum_{m\in\mathcal{M}}\phi_m\big)$ in this paper.



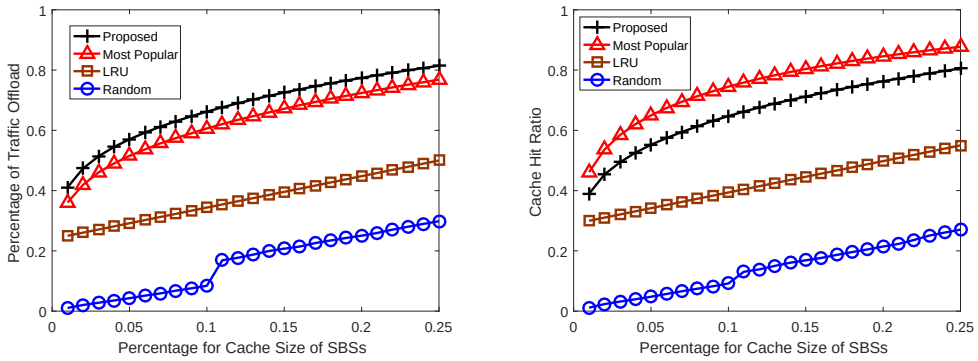
(a) Global content popularity

(b) Content size distribution

(c) Local content popularity

(d) Average overall arrival rate

Fig. 3. Different distributions in the IRCache trace and random setting.



(a) Percentage of traffic offload

(b) Cache hit ratio

Fig. 4. Percentage of traffic offload and cache hit ratio versus different caches sizes (percentage to the total content size) of each SBS.

Fig. 4 compares the performance of the proposed scheme with the baseline schemes in terms of percentage of traffic offload and cache hit ratio versus different caches sizes (percentage to the total content size) of each SBS. From Fig. 4, we can observe that as the cache size of each SBS increases, a higher percentage of traffic offload and higher cache hit ratio can be achieved in all the schemes. Most importantly, the proposed scheme can offload the most network traffic and the most popular caching scheme can achieve the highest cache hit ratio, which can be mathematically explained due to the difference between their optimization objectives. Besides, LRU caching scheme is inferior to both the proposed scheme and most popular caching scheme in both traffic offloading and cache hit ratio since LRU caching scheme only utilizes partial information of content popularity, while the random caching scheme has the worst performance since no information on content popularity is used. For instance, when the cache size of each SBS is set as 10% of the total size of contents, two observations can be made as follow: i) from Fig. 4(a), the proposed scheme, most popular caching scheme, LRU caching scheme, and random caching scheme can offload the network traffic by 64.7%, 59.0%, 34.3% and 8.5%, respectively; ii) from Fig. 4(b), these four schemes can achieve the cache hit ratio of 63.2%, 73.0%, 39.4% and 9.2%, respectively.

B. Numerical Results on Content Delivery

In this subsection, we evaluate by Monte-Carlo simulations the performance of our proposed joint user association and subcarrier-power allocation scheme for content delivery in the cache-enabled C-SCN over OFDMA downlinks. We consider that active users are uniformly distributed in the service area. Based on [25], we set the system bandwidth B as 2.5 MHz, subcarrier number N as 128, and carrier center frequency as 2.5 GHz. For the channel model, we set path loss exponent as 3.7, lognormal shadowing standard deviation as 8 dB, and noise power density as -174 dBm/Hz. The random channel fluctuations for small-scale fading are modeled as Rayleigh fading with unit average power. Based on the above performance of the proposed content placement scheme in terms of offloading network traffic and supporting content requests locally, we set the cache size of each SBS as 10% of the total size of contents. Each active user is set to have identical individual priority (i.e., $\omega_k = 1, \forall k \in \mathcal{K}$) and randomly requests only one of the considered contents in a time slot. Besides, we set the network priority as $\lambda \in \{1, 5, 10\}$, and consider that about 60% of the users requested contents are locally cached in the SBSs.

We set 128 Kbps as the required minimum data rate (i.e., $\{C_k^{\min}\}$) for delivering a content to a user. We average the performance over 100 random channel realizations to obtain the presented numerical results.

Note that according to the above settings, if the network priority λ takes the value of 1, then the proposed scheme is reduced to the general scheme for maximizing the total data rate under the same considered constraints in the cache-enabled C-SCN.

1) Convergence Performance of ADMM

Fig. 5 illustrates the convergence performance of the proposed ADMM in Algorithm 3 versus its complexity. For illustration purposes, we set the user number and the maximum transmit power $(K, p^{\max})$ as $(10, 25$ dBm$)$ and $(20, 30$ dBm$)$. Seen from Fig. 5, we can observe that in all the considered settings, Algorithm 3 requires at most 120 iterations to satisfy the given convergence condition. In particular, all the weighted sum of data rates (i.e., the considered optimization objective) decreases rapidly in [20, 90] iterations and then gradually converges. In addition, we can see that at the beginning of the iterative process of the proposed Algorithm 3, the values of weighted sum of data rates may be relatively large since the obtained transmit power matrix $\mathbf{P}$ only satisfies a part of the considered constraints, but Algorithm 3 always converges to a local optimum that satisfies all the considered constraints.

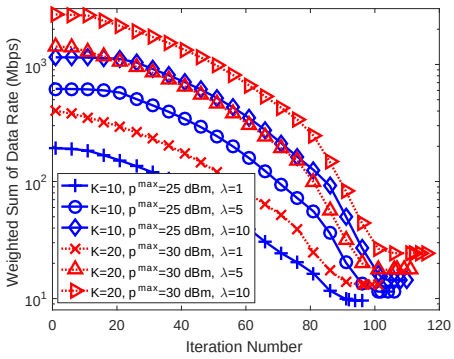


Fig. 5. Weighted sum of data rate versus iteration number.

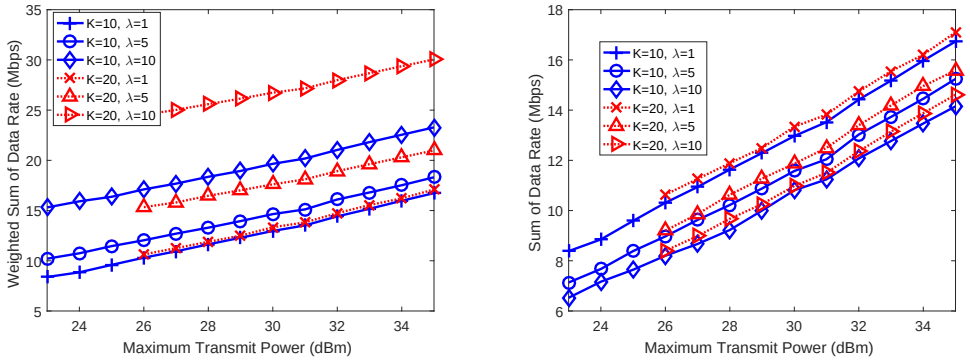
2) Effects of Different Maximum Transmit Power

Fig. 6 evaluates the effects of different maximum transmit power of SBSs on the weighted sum of data rates and sum of data rates in different settings. As seen from Fig. 6, with the increase of the maximum transmit power, all the achieved weighted sum of data rates and sum

of data rates also go up. In particular, a larger value of the chosen λ leads to larger achieved weighted sum of data rates but smaller sum of data rates, which can be explained by two facts: 1) mathematically, maximizing the objective $\lambda \sum_{k\in\mathcal{K}_1} \omega_k R_k + \sum_{k\in\mathcal{K}_0} \omega_k R_k$ is equivalent to maximizing $\sum_{k\in\mathcal{K}_1} \omega_k R_k + \frac{1}{\lambda} \sum_{k\in\mathcal{K}_0} \omega_k R_k$; 2) due to the larger network priority, more resources need to be allocated to the users whose requested contents are locally cached in the SBSs, while satisfying their required minimum data rates for content delivery. In addition, a larger number of users leads to an increase in the weighted sum of data rates and sum of data rates, as a result of the multiuser diversity gain.

3) Effects of Different User Numbers

Fig. 7 compares the weighted sum of data rates and sum of data rates versus different number of users in different settings. From Fig. 7, we can observe that all the weighted sum of data rates and sum of data rates go up with the increase of the number of users due to the multiuser diversity gain. Besides, a larger value of the chosen network priority λ leads to a larger weighted sum of data rates but a smaller sum of data rates as well.

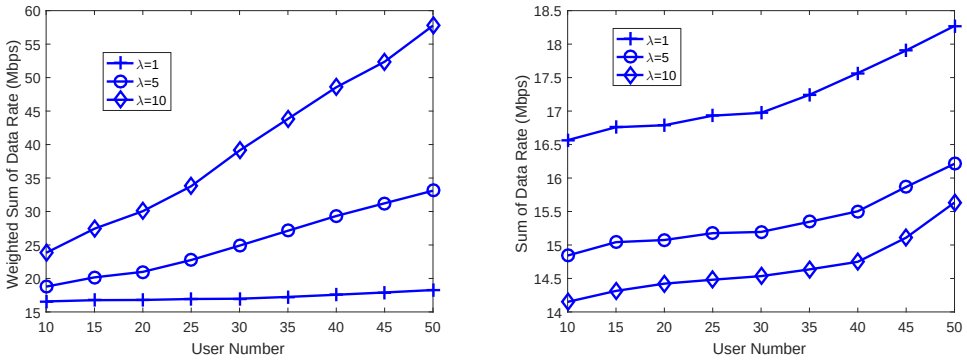


(a) Weighted sum of data rate (b) Sum of data rate

Fig. 6. Weighted sum of data rates and sum of data rates versus maximum transmit power.

VII. Conclusions

In this paper, we have proposed an efficient resource allocation framework for cache-enabled C-SCNs to achieve the benefits of content caching by considering two phases, i.e., content placement and content delivery. In particular, in the content placement phase, we have proposed a low-complexity distributed popularity-based framework for allocating cache sizes of SBSs to



(a) Weighted sum of data rate

(b) Sum of data rate

Fig. 7. Weighted sum of data rates and sum of data rates versus different numbers of users when $p^{\max} = 35$ dBm.

popular contents, aiming to maximize the expected sum of traffic offload in the network while satisfying content requests locally. Besides, in the content delivery phase, we have considered the wireless transmissions of contents from SBSs to users with given caching status in the network, and proposed a joint user association and subcarrier-power allocation scheme for min-rate guaranteed content delivery over OFDMA downlinks. To solve the formulated NP-hard optimization problem concerning the wireless resource allocation, we have proposed an approach using ADMM to decompose the problem into a series of simpler sub-problems for which optimal solutions can be easily obtained, and proposed the corresponding low-complexity algorithms to solve the sub-problems as well as the whole original problem, thereby realizing a design that is attractive for practical implementation. Numerical results from trace-based and Monte-Carlo simulations have been presented to illustrate the effectiveness of the proposed schemes in the cache-enabled C-SCNs.

REFERENCES

[1] X. Li, X. Wang, S. Xiao, and V. C. M. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," *in Proc. IEEE ICC*, pp. 5652-5657, Jun. 2015.

[2] R. Wang, X. Peng, J. Zhang, and K.B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77-83, Aug. 2016.

[3] H. Zhou, H. Wang, X. Li, and V. C. M. Leung, "A survey on mobile data offloading technologies," *IEEE Access*, vol. 6, pp. 5101-5111, Jan. 2018.

[4] X. Li, X. Wang, K. Li, H. Chi, and V. C. M. Leung, "Resource allocation for content delivery in cache-enabled OFDMA small cell networks," *in Proc. IEEE VTC-Fall*, pp. 1-5, Sept. 2017.

[5] H. Zhou, H. Zheng, J. Wu, and J. Chen, "Energy efficiency and contact opportunities trade-offs in opportunistic mobile networks", *IEEE Trans. Vehi. Tech.*, vol. 65, no. 5, pp. 3723-3734, May 2016.

[6] X. Li, P. Wu, X. Wang, K. Li, Z. Han, and V. C. M. Leung, "Collaborative hierarchical caching in cloud radio access networks," *in Proc. IEEE INFOCOM Workshops*, May 2017.

[7] X. Ge, X. Li, H. Jin, J. Cheng, and V. C. M. Leung, "Joint user association and scheduling for load balancing in heterogeneous networks," *in Proc. IEEE GLOBECOM*, pp. 1-6, Dec. 2016.

[8] X. Ge, X. Li, H. Jin, J. Cheng, and V. C. M. Leung, "Joint user association and user scheduling for load balancing in heterogeneous networks," *IEEE Trans. Wireless Commun.*, Feb. 2018. DOI: 10.1109/TWC.2018.2808488

[9] H. Zhou, V. C. M. Leung, C. Zhu, S. Xu, and J. Fan, "Predicting temporal social contact patterns for data forwarding in opportunistic mobile networks", *IEEE Trans. Vehi. Tech.*, vol. PP, no. 99, 2017.

[10] X. Li, X. Wang, K. Li, Z. Han, and V. C.M. Leung, "Collaborative multi-tier caching in heterogeneous networks: modeling, analysis, and design", *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926-6939, Oct. 2017.

[11] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," *in Proc. IEEE ICC*, pp. 1-6, May 2016.

[12] X. Li, X. Wang, K. Li, and V. C. M. Leung, "Collaborative hierarchical caching for traffic offloading in heterogeneous networks," *in Proc. IEEE ICC*, May 2017.

[13] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: wireless video content delivery through distributed caching helpers," *in Proc. IEEE INFOCOM*, Mar. 2012.

[14] X. Li, X. Wang, K. Li, and V. C. M. Leung, "CaaS: caching as a service for 5G networks," *IEEE Access*, vol. 5, pp. 5982-5993, May 2017.

[15] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," *in Proc. IEEE GLOBECOM*, pp. 1-6, Dec. 2015.

[16] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: an energy-efficient approach to improve Quality of Service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207-1221, May 2016.

[17] J. Li, H. Chen, Y. Chen, Z, et al., "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115-2129, Aug. 2016.

[18] Q. Chen, F.R. Yu, T. Huang, R. Xie, J. Liu, and Y. Liu, "Joint resource allocation for software defined networking, caching and computing," *in Proc. IEEE GLOBECOM*, pp. 1-6, Dec. 2016.

[19] Y. Jin, Y. Wen, and C. Westphal, "Towards joint resource allocation and routing to optimize video distribution over future Internet," *in Proc. IFIP Networking*, pp. 1-9, May 2015.

[20] A. Liu and V. K. N. Liu, "Exploiting base station caching in MIMO cellular networks: opportunistic cooperation for video streaming," *IEEE Trans. Signal Processing*, vol. 63, no. 1, pp. 57-69, Jan. 2015.

[21] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118-6131, Sept. 2016.

[22] R.G. Stephen, and R. Zhang, "Green OFDMA resource allocation in cache-enabled CRAN," *in Proc. IEEE OnlineGreenComm*, Dec. 2016.

[23] S. Boyd, N. Parikh, et al., "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp.1-122, 2011.

[24] C. Shen, T. H. Chang, K. Y. Wang, Z. Qiu, and C. Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect csi: An admm approach," *IEEE Trans. on Signal Processing*, vol. 60, no. 6, pp. 2988-3003, Jun. 2012.

[25] X. Li, X. Ge, X. Wang, J. Cheng, and V. C. M. Leung, "Energy efficiency optimization: joint antenna-subcarrier-power allocation in OFDM-DASs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7470-7483, Nov. 2016.

[26] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5G cloud radio access networks," *IEEE Network*, vol. 31, no.4, pp. 35-41, July 2017.

[27] X. Li, X. Wang and V. C. M. Leung, "Optimizing power allocation in wireless networks: are the implicit constraints really redundant?" *Computer Communications*, vol. 111, pp. 153-164, Oct. 2017.

[28] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan. *Mixed-Integer Nonlinear Optimization*. Argonne National Laboratory, 2012.

[29] S. Boyd and L. Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004.

[30] M. Hong, Z. Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," Oct. 2014. http://arxiv.org/abs/1410.1390

[31] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," Apr. 2018. https://arxiv.org/abs/1511.06324

[32] National Laboratory for Applied Network Research, Weekly Squid HTTP Access Logs, http://www.ircache.net/.

[33] A. B. Downey, "The structural cause of file size distributions," *in Proc. IEEE MASCOTS*, pp. 361-370, Aug. 2001.

[34] M. Chrobak and J. Noga, "LRU is better than FIFO," *in Proc. ACMSIAM SODA*, pp. 78?1, Jan. 1998.